DOCUMENT RESUME

ED 444 999 TM 031 535

AUTHOR Tay-Lim, Brenda Siok-Hoon; Harwell, Michael
TITLE Effects of Number of Items and Examinees on Parameter Estimation in Item Response Theory: A Research Synthesis.
PUB DATE 1997-03-00
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); Item Response Theory; *Literature Reviews; Research Utilization; *Synthesis
IDENTIFIERS *Item Parameters

ABSTRACT

The purpose of this paper is to illustrate how a quantitative literature review or research synthesis can be used to aggregate information in the item response theory literature to address specific research questions. The research questions focused on the magnitude of the contribution of the number of items and number of examinees to the accuracy of parameter estimates in the two-parameter item response theory model. The results from 7 studies used in the research synthesis suggest that more than 63% of the variables in a commonly used indicator of parameter estimation accuracy was attributable to these 2 factors, and that this result holds across a variety of factors. Other questions that a research synthesis might be used to address in the item response theory literature are described. (Contains 13 tables and 31 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 444 999

Effects of Number of Items and Examinees on Parameter Estimation in Item Response Theory: A Research Synthesis

Brenda Siok-Hoon Tay-Lim
Michael Harwell

University of Pittsburgh

B. Tay-Lim

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM031535

Paper presented at the annual meeting of the American Educational Research Association, Chicago, March, 1997.

BEST COPY AVAILABLE

Abstract

The purpose of this paper was to illustrate how a quantitative literature review or research synthesis can be used to aggregate information in the item response theory literature to address specific research questions. The research question focused on the magnitude of the contribution of the number of items and number of examinees to the accuracy of parameter estimates in the two-parameter item response theory model. The results suggested that more than sixty-percent of the variation in a commonly used indicator of parameter estimation accuracy was attributable to these two factors, and that this result holds across a variety of factors. Other questions that a research synthesis might be used to address in the item response theory literature are described.

The presence of a considerable body of empirical literature in item response theory (IRT) provides an opportunity to aggregate this information to address important research questions and to identify gaps in this literature. This paper outlines how a quantitative literature review or research synthesis can be used for this purpose. A research synthesis attempts to quantitatively aggregate the findings from independent studies of the same phenomenon. Study findings are captured through effect sizes, for example, the standardized mean difference between a treatment and control group. Variation among effect sizes is then statistically modeled as a function of various predictor variables believed to be related to the effect sizes (e.g., number of subjects in a study, methodological quality of a study). The strength of research synthesis lies in its ability to examine information accrued over multiple studies and several operational definitions. Ideally, a research synthesis allows estimates of magnitudes of effects for theorized relations to be aggregated in ways that rarely would or could be tested within one study.

Most research syntheses are performed in non-quantitative research settings, for example, studying the relationship between gender and performance on standardized mathematics tests (e.g., Hyde, Fennema, & Lammon, 1990). Harwell (1992), Harwell, Rubinstein, Hayes, and Olds (1992), Lix, Keselman, and Keselman (1997) and others have extended this methodology to aggregating the results of computer simulation studies in statistics; this paper proposes that methods of research synthesis be extended to the IRT computer simulation literature. Thus, the purpose of the paper is to describe how a research synthesis might be used in this arena to provide evidence about important questions. The ability of a research synthesis to point out gaps in a literature and thus suggest future studies is also emphasized. Hopefully, this framework will encourage measurement specialists and practitioners to consider employing this methodology.

Description of the problem

Previous attempts to synthesize the IRT simulation literature have been narrative in nature. For example, Baker (1992) narratively summarized simulation studies on marginal Bayesian parameter estimation, Hambleton and Swaminathan (1985) summarized the model-data fit empirical literature, and Hulin, Drasgow, and Parsons (1983) summarized empirical literature pertaining to parameter estimation in IRT models. Of course, narrative (qualitative) summaries appear routinely in published (primary) simulation studies in IRT. Narrative

reviews of IRT simulation studies have much to recommend them, but can be expected to possess the incumbent difficulties of such reviews. These include a lack of an overarching theory to guide their interpretation, which may limit the interpretation of the findings to the conditions modeled, and the impressionistic nature of these studies (Harwell, 1992). These shortcomings can be addressed by employing a quantitative literature review (i.e., a research synthesis).

We focus on a particular setting to illustrate a research synthesis in IRT without making claims of its importance: A frequent concern in test calibration in IRT is how many items and examinees are needed to produce parameter estimates of the desired accuracy. Narrative reviews have documented that the estimation of item and ability parameters tends to be more accurate when there are larger numbers of items and examinees. However, in many practical testing situations there are constraints on the number of items that can be administered and the number of available examinees.

The role of factors like test length and number of examinees in parameter estimation has been heavily researched for dichotomously-scored IRT models using computer simulation studies, and less well-researched for other IRT models (e.g., polytomous models). These simulation studies may report dozens, hundreds, or even thousands of outcomes reflecting the accuracy of parameter estimation for various test lengths, examinee sample sizes, IRT models, methods of calibration, prior distributions of parameters, etc. The sheer volume of information and its complexity can make it difficult to provide relatively precise answers to questions such as the number of items and examinees required to produce accurate parameter estimates. This literature is also capable of yielding contradictory findings, adding to the difficulty of summarizing information.

Review of the IRT Simulation Literature on the Accuracy of Parameter Estimation

To set the stage for the research synthesis, we begin with a brief narrative review of the IRT literature investigating the accuracy of parameter estimation for dichotomous response models. Studies have investigated the effect of varying estimation methods, prior distributions of ability and item parameters, ability distributions, and numbers of replications (e.g., Drasgow, 1982; Gifford & Swaminathan, 1990; Harwell & Janosky, 1991; Hulin, Lissak, & Drasgow, 1982; Kim, Cohen, Baker, Subkoviak, & Leonard, 1994; Lim & Drasgow, 1990; Seong, 1990; Skaggs & Stevenson, 1989; Stone, 1992; Swaminathan & Gifford, 1986; Yen, 1987).

Unambiguous conclusions about the magnitude of the contribution of these factors are rare; instead, the findings depend on the conditions modeled.

For example, it is generally agreed that test length and examinee sample size affect the accuracy of estimation, but there is less agreement on what test lengths and examinee sample sizes are needed to achieve a specified level of accuracy for a given IRT model, ability distribution, etc. For example, the role of test length and number of examinees in minimizing estimation error is not the same for the 1-, 2-, and 3-parameter IRT models. Previous studies have indicated that, other things being equal, 1- and 2-parameter IRT models require fewer examinees and items than a 3-parameter IRT model for accurate parameter estimates to be obtained.

Lord (1968) suggested that a sample size greater than 1,000 examinees and more than 50 items are needed for adequate estimates in a 3-parameter IRT model. Swaminathan and Gifford (1986) found that item discrimination parameters were poorly estimated when sample size was small (50, 200) and test length was short (10, 15, 20). Ree and Jensen (1980) also found sample size requirements for item parameter estimation to be substantial for the 3-parameter IRT model (Hulin, Lissak, & Drasgow, 1982). These studies typically use the square root of the average squared difference between a parameter estimate and the true parameter (root mean squared deviation or RMSD) or the correlation between estimated and true parameters as indicators of estimation accuracy. These indicators are known as effect sizes in a research synthesis.

Hulin, et al. (1982) found that tests of as few as 30 items combined with sample sizes of 500 examinees for a 2-parameter model or 1,000 for the 3-parameter model were sufficient for accurate parameter estimation. Other studies have found that sample size and test length requirements for the 3-parameter model can be relaxed if an informative prior is imposed on the item discrimination and guessing parameters (Drasgow, 1989; Gifford & Swaminathan, 1990; Harwell & Janosky, 1991; Seong, 1990; Skaggs & Stevenson, 1989; Swaminathan & Gifford, 1986). In general, the simulation literature suggests that Bayesian procedures like those described in Mislevy (1986) produce more accurate estimates than the non-Bayesian joint maximum likelihood estimation (JMLE) procedure when the sample size is small and test length is short in the 2- and 3-parameter IRT models (Skaggs & Stevenson, 1989; Yen, 1987).

Still, the availability of information about the contribution of test length and examinee sample size to accurate parameter estimation has not yet produced precise conclusions about the magnitude of the contribution of

these (or other) factors for various IRT models. Perhaps this is attributable to the impressionistic nature of these studies, or perhaps to the absence of an overarching framework to guide the interpretation of these studies. We use a research synthesis to try to provide such information.

The synthesis is organized using the format recommended by Cooper (1982) (1) Problem Formulation (2) Data Collection (3) Data Evaluation (4) Data Analyses and Interpretation (5) Presentation of Results.

Problem Formulation

The starting point of a research synthesis is specification of one or more research questions. Our research question is simply: What is the magnitude of the contribution of the number of examinees and test length on the accuracy of estimated item and ability parameters for varying IRT models, estimation methods, types of prior distributions, types of ability distributions, and numbers of replications? Ideally, the results will complement existing narrative reviews of the IRT literature related to this topic and provide guidance for measurement practitioners and specialists.

Data Collection

The population of studies targeted included in this synthesis examined at least one of the selected effect size measures, the RMSD or the correlation of true and estimated parameters. These indicators of estimation accuracy appear regularly in the IRT computer simulation literature. In addition to the number of examinees and items, information was collected on the following variables: type of IRT model, method of parameter estimation, presence or absence of prior distributions for item and ability parameters, shape of the ability distribution, and the number of replications employed.

Initially, 10 studies were obtained as part of a research synthesis for a class project. Subsequent studies were identified by a systematic search of the Educational Resources Information Center (ERIC) database for the years 1982 to 1995. The keywords 'Item Response Theory' and 'Item Parameter Estimation' were used to conduct the search, which provided 17 additional studies. Nine of the 17 studies were conference papers and ETS technical reports which could not be accessed in time to complete this study. The reference lists from those

articles which were included yielded 2 additional articles. A total sample of of 20 papers were accessed through this process.

Seven of the 20 studies were eliminated because they were theoretical papers and did not have relevant information (de Gruijter, 1985; Hambleton & Jones, 1994; Harwell, et al., 1988; Harwell & Baker, 1991; Mislevy & Sheehan, 1989; Stocking, 1990; Tsutakawa & Johnson, 1990). In addition to these 7 studies, 6 additional studies were excluded because of a failure to report information needed to calculate effect sizes. One study was excluded because it reported the effect size in terms of item characteristics curves (Drasgow, 1989). Another was excluded because it did not directly address the issue of accuracy of the parameter estimation (Baker, 1990). One other study reported data in the form of graphs, making it impossible to accurately extract the necessary data (Mislevy & Stocking, 1989). A fourth study (Ackerman & Stone, 1992) was excluded since it used a graded response model in the study. Two studies could not be coded in a manner that was consistent with the coding scheme developed for the research synthesis (Kim, Cohen, Baker, Subkoviak, & Leonard 1994; Kim & Nicewander, 1993). Thus, 7 articles yielding a total of 119 effect sizes were included in the research synthesis and are given in the reference list.

Strictly speaking, the 7 studies used in the research synthesis represented a convenience sample that may differ from the targeted population of studies. However, there is little reason to believe that the sample of studies is radically different from those in the population.

Study Features. The 7 studies used in the research synthesis had several common elements. Three of the studies examined the accuracy of parameter estimates for the joint maximum likelihood estimation procedure (JMLE) procedure and the marginal maximum likelihood estimation (MMLE) procedure (Skaggs & Stevenson, 1989; Swaminathan & Gifford, 1986; Yen, 1987), and 3 studies examined the effect of prior distributions on the parameter estimates (Gifford & Swaminathan, 1990; Harwell & Janosky, 1991; Seong, 1990). All of the studies varied the number of examinees and number of items as part of investigating the accuracy of parameter estimation. The sample sizes used in the 7 studies ranged from 25 to 2000, with sample sizes of 200, 500, 1000, 2000 modeled in Hulin, et al., (1982), 100, 200, 400 in Swaminathan and Gifford (1986), 1000 in Yen (1987), 500 and 2000 in Skaggs and Stevenson (1989), 25, 50, 100, 150, 200, 500 in Gifford and Swaminathan (1990), 100 and 1000 in Seong (1990), and 75, 100, 150, 250, 500, 1000 in Harwell and Janosky (1991). The number of

items used in the 7 studies ranged from 10 to 60, with 15, 30, 60 items modeled in Hulin, et al., (1982), 25 and 35 items in Swaminathan and Gifford (1986), 10, 20, 40 items in Yen (1987), 15 and 35 items in Skaggs and Stevenson (1989), 15, 25, 35, 50 items in Gifford and Swaminathan (1990), 45 items in Seong (1990), and 15 and 25 items in Harwell and Janosky (1991).

The number of replications ranged from 1 to 20. Some studies used only 1 replication (Harwell & Janosky, 1991; Swaminathan & Gifford, 1986; Yen, 1987), while others used 2, 4, 5 or 20 replications. Some authors used the 3-parameter IRT model (Gifford & Swaminathan, 1990; Skaggs & Stevenson, 1989; Swaminathan & Gifford, 1986; Yen, 1987), some the two-parameter model (Harwell & Janosky, 1991; Seong, 1990), and some used both 2- and 3-parameter models (Gifford & Swaminanthan, 1990; Hulin, et al., 1982). All studies except Hulin, et al. (1982) employed pseudo-Bayesian or Bayesian procedures for parameter estimation, so prior distributions of some sort were used. Finally, Gifford and Swaminathan (1990), Seong (1990) and Yen (1987) generated item responses from normal and nonnormal distributions, while the remaining studies modeled only normally-distributed data. Tabulating the conditions modeled in these studies may identify gaps in this literature and suggest future studies.

Coding. A total of 16 variables were initially coded for each study; however, because of deficient reporting in the studies (described below) or because there was no variation in a coded variable across studies, 7 of the 16 variables were dropped. The remaining 9 coded variables are listed in Table 1. All studies were independently coded by two investigators and all discrepancies in coding were resolved by mutual agreement among the coders. The reliability indices are illustrated by reporting these results for the 2-parameter logistic model in Table 2. For the categorical value variables, Cohen's kappa was used, and for the continuous values variables an interclass correlation was used. The values in Table 2 suggests that the coding was reliable.

Two of the variables were coded to be dichotomous because of a lack of detailed information in primary studies. Type of prior distribution was coded as "0" when a prior distribution was not used or when a prior was symmetric, and "1" when a prior distribution was skewed. Ideally, prior distributions would have been coded as a function of both shape (e.g., symmetric, skewed) and variance (smaller variances implying a more informative prior) but this was not possible because of a lack of information about priors in primary studies. For similar reasons, the ability distribution variable was also dichotomized, with a "0" used for distributions that were

symmetric, and a "1" for distributions that were skewed. For example, all reported normally-distributed ability distributions were known to have a skewness and kurtosis of 0, whereas the skewed distributions that were modeled typically reported little specific information about skewness and kurtosis.

The coding of the effect sizes deserves special note because of the occasional difficulties. For example, Gifford and Swaminathan (1990) and Swaminathan and Gifford (1986) reported mean square differences for estimated parameters; therefore, the RMSD could be calculated directly. On the other hand, Skaggs and Steveson (1989) reported the correlation between estimated and true item parameters by ability groups. In this case, a weighted average of the correlations across the low, medium and high ability groups was calculated. Similarly, Seong (1990) reported the RMSD by ability groups, and again an average was computed across ability groups. Also, the 7 studies produced multiple effect sizes, but this was not a problem because the effect sizes were based on independently generated data.

Another aspect of calculating the effect sizes was their use in hypothesis-testing later in the research synthesis. For hypothesis-testing purposes, the RMSDs and correlations needed to be normally-distributed. Since a distribution of RMSDs is usually positively skewed, a log transformation was applied under the assumption that the quantities used to computed the RMSDs (i.e., estimated parameters) were themselves normally-distributed. If this holds, the log-transformed RMSD variable will be asymptotically normally-distributed with known mean and variance (Kendall, & Stuart, 1942). A Fisher r-to-z transformation was applied to the correlations to produce an approximate normal distribution for the transformed values (Cooper & Hedges, 1994).

Missing Data. Several studies did not have information on all coded variables and a code of "missing" was used in such cases. Generally, there were two kinds of missing data. For example, Swaminathan and Gifford (1986) did not report the RMSD and correlation values for the 25 item, 400 examinees case because of non-convergence in the JMLE procedure. Another kind of missing data occured when studies only reported one of the effect sizes rather than both. For example, Hulin, et al. (1982) and Skaggs and Steveson (1989) reported the correlation between the true and estimated parameters for item discrimination and difficulty parameters, but not the RMSDs, whereas Seong (1990) reported RMSDs for estimated and true parameters but not the correlations.

Results

Results for various descriptive analyses of the effect sizes are presented first, followed by inferential analyses composed of between- and within-study tests of heterogeneity and a regression analysis. The coded simulation factors, such as number of items and examinees, were treated as predictor variables and the effect sizes as criterion variables.

Correlations Among Variables

Correlations among the coded variables are reported in Tables 3-5. Note that interpretation of the statistically significant correlations is enhanced if they are squared to reflect explained variation.

Correlations Among Predictors. There were not many surprising results in the correlations among the predictors for item and ability parameters. Naturally, the correlations among type of IRT model, method of parameter estimation, and prior distribution were larger. And there is little reason to be surprised by the small correlations between these three variables and other predictors (e.g., number of items, number of examinees, and number of replications).

Correlations Between Predictor and Criterion Variables. More interesting patterns emerged in the correlations between the predictor and criterion variables. Generally, correlations between log-RMSDs were higher (though still quite small) for more complicated IRT models and for the JMLE procedure and smaller examinee samples. The association between log-RMSDs and more complicated models is reasonable considering that the accuracy of estimation often decreases as the number of estimated parameters per model increases. For example, the relationship implies that item discrimination parameters are slightly more accurately estimated as estimation procedures shifted from a JMLE to a MMLE or fully Bayesian estimation procedure. Similarly, as the number of examinees increases, the accuracy of estimation increases slightly, resulting in a smaller log-RMSD.

However, the correlations between log-RMSD and other factors were generally low. Consider Tables 3 and 4 which focus on item parameters. Of course, as the number of examinees increased log-RMSD values tended to decrease, but the relationship was moderately weak ($r = -.29$). Similarly, the correlations between log-RMSD and prior distribution and number of replications was small. The largest reported correlation in Table 3 (discrimination parameters) for log-RMSD was with estimation method

($r = -.36$), whereas this same correlation for difficulty parameters in Table 4 failed to materialize. This makes sense because estimation of discrimination parameters has been more difficult (and, hence, more method-dependent) than estimation of difficulty parameters. Still, the explained variation of $(-.36^2) = 13\%$ seems low.

Significant relationships were also found for ability distribution and the number of examinees. As ability distributions shifted from normal to non-normal, the log-RMSDs tended to decrease. The log-RMSD for item difficulty also tended to decrease with increases in the number of examinees. Use of a more complicated IRT model tended to decrease the accuracy of ability estimation, whereas improved accuracy of ability estimation tended to occur in shifting from a JMLE to a MMLE or Bayesian procedure. The increased accuracy of ability estimates with longer tests ($r = -.69$) is logical since ability will be more precisely estimated as test length increases. There was also a significant negative relationship between the two criterion variables, which is reasonable since decreases in RMSD suggest more accurate estimation, which should be associated with higher correlations between estimated and true parameters.

Homogeneity Tests

Homogeneity tests were used to further investigate the relationship between the effect sizes and predictor variables. The initial homogeneity test was done to simply test whether the effect sizes were homogeneous (see Alexander, Scozzaro, & Borodkin, 1989 for a description of tests of homogeneity); if they were heterogeneous the next step would be to try to pinpoint the sources of heterogeneity. Only results for the 2-parameter case are reported since most of the effect sizes were associated with this model.

The homogeneity tests were all statistically significant at the $\alpha = .05$ level for both the log-RMSD and Fisher r-to-z statistics. The tests for log-RMSDs were $\chi^2 = 1776$ (df=72) for the discrimination parameters, $\chi^2 = 790.2$ (df=86) for difficulty parameters, and $\chi^2 = 2223.8$ (df=65) for ability parameters. For the Fisher r-to-z statistics, the tests were also significant at $\alpha = .05$ for discrimination (χ^2 = is 552, df=94), difficulty ($\chi^2 = 1118$, df=101), and ability parameters ($\chi^2 = 1890.3$, df=53). These results suggested that the effect sizes were not homogenous across the various combinations of factors (e.g., presence or absence of a prior distribution).

Between- and Within-Study Analyses

Next, homogeneity tests were done for the effect sizes both between- and within-studies to try to pinpoint the sources of the heterogeneity. These results are reported in Tables 6-11. Consider the homogeneity tests using the log-RMSD for item discrimination parameters in Table 6. The 5 studies examined differed statistically in average log-RMSD values, with the Swaminathan and Gifford (1986) study showing the largest average log-RMSD value. The log-RMSD values were also found to be heterogeneous within each of the 5 studies, with the Swaminathan and Gifford (1986) study again showing the most heterogeneity. The large differences in log-RMSDs within this study were the result of using both JMLE and Bayesian estimation procedures in parameter estimation in the 3-parameter model. These results indicated that the conditions modeled in this study produced parameter estimates of significantly different accuracy.

It is equally important to consider what the lack of within-study heterogeneity in the effect sizes in the Hulin, et al., and Skaggs and Stevenson (1989) studies implies. This result in Table 6 means that the conditions modeled produced (statistically) the same estimation accuracy for discrimination parameters. Interestingly, a few

of the conditions modeled in these studies were the same as those modeled in studies showing significant heterogeneity. This may be the result of a Type I or II error in hypothesis testing, or perhaps can be explained by the way the simulation was done. At the least, it is a difference worth pursuing to better reconcile these findings. Similar patterns were found for the log-RMSD values for difficulty and ability parameters (Tables 7-8).

The 7 studies differed significantly in the average correlation between true and estimated discrimination parameters (Table 9). The Hulin, et al., Gifford and Swaminathan, and Harwell and Janosky studies also showed significant within-study heterogeneity for discrimination parameters. For the Hulin, et al. study the heterogeneity was attributable to the use of different IRT models (2- vs. 3-parameter; for the Gifford and Swaminathan study the heterogeneity was attributable to the use of normal and non-normal ability distributions; for the Harwell and Janosky study the heterogeneity was attributable to the presence or absence of small prior distribution variances. Similar patterns arose for difficulty and ability parameters (Tables 10-11).

General Linear model

Regression analyses allow a combination of continuous and dichotomous variables to be used to explore the relationship between the effect sizes and the predictors. Factors treated as nominal variables, such as type of IRT model and method of parameter estimation, were dummy-coded prior to inclusion in the regression model. Variables which were metric were introduced unchanged into the regression model. The Proc REG procedure in SAS (SAS Institute, 1990) was used to perform a weighted least squares regression in which the weights were the inverse of the variances of the log-RMSDs and r-to-z statistics. Preliminary analyses suggested that the prior distribution and method of parameter estimation variables were giving the same information, and the prior distribution variable was dropped.

Given the main research question, the focus of the regression analysis was on the contribution of the number of examinees and items predictors to explaining variation in the log-RMSD and Fisher r-to-z statistics. The variation accounted for by the two predictors, with other predictors held constant, is reported in T ables 12 and 13. For the log-RMSD variable (Table 12) the explained variation for discrimination parameters was approximately 62%; for difficulty parameters it was 63%; for ability parameters it was 63%. Thus, the variation in log-RMSD values for discrimination, difficulty, and ability parameters explained by the number of examinees

and number of items, with other predictors held constant, was substantial, accounting for at least 76% of the variation explained by the full regression model. The R^2 values for the full model and for the contribution of number of items and examinees to the full model are lower for the Fisher r-to-z statistics (see Table 13).

With appropropriate caution, it would also be possible to extrapolate these findings by plugging in predictor values not present in the sample data, to the fitted model to obtained predicted log-RMSD and correlation values. This could, for example, help to identify the point(s) at which further increases in estimation accuracy are marginal for increased numbers of examinees.

Discussion

The purpose of this paper was to illustrate how a quantitative literature review or research synthesis can be used to aggregate information in the item response theory literature to address specific research questions. To illustrate the potential of research synthesis in this setting, the relationship between parameter estimation accuracy and number of items and examinees was studied. The results of a research synthesis suggested that more than sixty percent of the variation in a commonly used indicator of estimation accuracy was attributable to the number of items and number of examinees. While the lion's share of concern over estimation accuracy should continue to focus on the number of items and examinees, the research synthesis results suggest that other factors, such as the method used to estimate parameters, play important roles in estimation accuracy (i.e., account for a non-negligible amount of explained variance).

Research synthesis also has several limitations. One of the most important is the representativeness of the sample of accessed studies. Another limitation of the reported research synthesis was the inability to distinguish between estimation bias and the error variance associated with the parameter estimates. In replicated studies these two sources of variation can be distinguished; in unreplicated studies they cannot (Gifford& Swaminathan, 1990). The presence of replicated studies in a research synthesis alows model misspecification (i.e., whether all of the predictors variables needed to explain variation in the effect sizes are in the model) to be investigated. Increased use of replications in simulation studies in item response theory will allow this important facet of a research synthesis to be exploited.

What important questions could a research synthesis be applied to in item response theory? Examples include (1) What is the increase in parameter estimation accuracy as item response models increase in complexity? (2) How do various estimation methods compare in estimation accuracy? (3) Which method of detecting multidimensionality has the best statistical properties? (4) Which method of detecting differential item functioning has the best statistical properties? Primary studies of these topics exist in the item response literature; what is needed is a mechanism which allows their findings to be quantitatively aggregated in ways that complement narrative reviews. A research synthesis represents one such vehicle.

References

Ackerman, R. & Stone, C.A. (1992). A monte carlo study of marginal maximum likelihood parameter estimates for the graded model. Paper presented at the meeting of the National Council on Measurement in Education,

Alexander, R.A., Scozzaro, M.J., & Borodkin, L.J. (1989). Statistical and empirical examination of the chi-square test of homogeneity of correlations in meta-analysis. Psychological Bulletin, 106, 329-331.

Baker, F.B. (1990). Some observations on the metric of PC-BILOG results. Applied Psychological Measurement, 14, 139-150.

Baker, F. B. (1992). Item response theory. New York: Marcel Dekker.

Cooper, H.M. (1982). Scientific guidelines for conducting integrative literature reviews. Review of Educational Research, 52, 291-302.

Cooper, H., & Hedges, L. V. (1994). The Handbook of Research Synthesis. New York, NY. Russell Sage Foundation.

De Gruijter, D.N.M. (1985). A note on the asymptotic variance-covariance matrix of item parameter estimates in the Rasch model. Psychometrika, 50, 247-249.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. Applied Psychological Measurement, 13, 77-90.

*Gifford, J.A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters in item response models. Applied Psychological Measurement, 14, 33-43.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

Harwell, M. (1992). Summarizing Monte Carlo results in methodological research. Journal of Educational Statistics, 17, 297-313.

Harwell, M.R., & Baker, F.B. (1991). The use of prior distributions in marginalized Bayesian parameter item parameter estimation: A didactic. Applied Psychological Measurement, 15, 375-389.

Harwell, M.R., Baker, F.B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. Journal of Educational Statistics, 13, 243-271.

Harwell, M.R., Rubinstein, E.N., Hayes, W., & Olds, C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17, 315-339.

*Harwell, M.R., & Janosky, J.E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. Applied Psychological Measurement, 15, 279-291.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory: Application to psychological measurement. Homewood, Illinois: Dow Jones-Irwin.

*Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: a monte carlo study. Applied Psychological Measurement, 6, 249-260.

Hyde, J.S., Fennema, E., & Lammon, S.H. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107, 139-155.

Kim, J.K., & Nicewander, W.A. (1993). Ability estimation for conventional tests. Psychometrica, 58, 587-599.

Kim, S.H., Cohen, A.S., Baker, F.B., Subkoviak, M.J., & Leonard, T. (1994). An investigation of the hierarchical Bayes procedures in item response theory. Psychometrika, 59, 405-421.

Lix, L.M., Keselman, J.C., & Keselman, H.J. (1997). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Review of Educational Research, 66, 579-619.

Lord, F.M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.

Mislevy, R.J. & Stocking, M.L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

SAS Institute. (1990). SAS/STAT User's Guide. Cary, NC.

*Seong, T.J. (1990). Sensitivity of marginal maximal likelihood estimation of item and ability parameters to characteristics of the prior ability distributions. Applied Psychological Measurement, 14, 299-311.

*Skaggs, G. & Stevenson, J. (1989). A comparison of pseudo-bayesian and joint maximal likelihood procedures for estimating item parameters in the three-parameter IRT model. Applied Psychological Measurement, 13, 391-402.

Stone, C.A. (1992). Recovery of marginal maximal likelihood estimates in the two-parameter logistic response model: an evaluation of MULTILOG. Applied Psychological Measurement, 16, 1-16.

*Swaminathan, H. & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. Psychometrica, 51, 589-601.

*Yen, W. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

* Used in the research synthesis

Table 1

Coded Variables Use in the Analyses

- Item response model: 1=one-, 2=two-, 3=3-parameter dichotomous model
- Method of item estimation: 1=JMLE, 2=MMLE/Pseudo-Bayesian, 3=Bayesian
- Distribution of parameters used to generate the item responses: 0=symmetric, 1=skewed
- Type of prior distribution specified for the parameter estimation: 0=symmetric, 1=skewed
- Number of examinees: range= 25,2000
- Number of items: range=10,60
- Number of replications: range=1,20
- RMSD between true and estimated item and ability parameters
- Correlation between true and estimated item and ability parameters

Table 2

Inter-Rater Reliability Estimates for Coding of Variables Included in Analyses

Variables	Reliability		
	item discrimination	item difficulty	ability
Study identification number[a]	1.000	1.000	1.000
Type of IRT Model[a]	1.000	1.000	1.000
Method of Parameter Estimation[a]	1.000	1.000	1.000
Prior distribution[a]	1.000	.631	.618
Ability distribution[a]	1.000	1.000	1.000
Number of items[b]	.853	.844	.753
Number of Examinees[b]	.900	.924	.988
Number of Replications[b]	.943	.823	.795
RMSD[b]	.875	.216	.912
Correlation[b]	.841	.625	.867

a - Based on Cohen's Kappa K=(Po-Pe)/(1-Pe)

b - Based on interclass correlation (Design 3:mixed effect model) r=(BMS-EMS)/(BMS+EMS)

Table 3

Correlation Between Predictors and Criterion Variables for Item Discrimination Parameter

	Model	Method	Prior Dist.	Ability Dist.	Number of items	Number of exam.	Number of replica	RMSD	Correlation
IRT Model	1.000								
Method of Parameter Estimation	**-.448**[a] n = 105 p<.001	1.000							
Prior Distribution	**-.194**[a] 103 p=.049	**.678**[a] 103 p<.0001	1.000						
Ability Distribution	**.187**[a] 105 p=.056	**.080**[a] 105 p=.714	**.092** 103 p=.351	1.000					
Number of Items	**.018**[b] 105 p=.169	**-.075**[b] 105 p=.019	**.001**[b] 103 p=.758	**.047**[b] 105 p=.027	1.000				
Number of Examinees	**.076**[b] 105 p=.004	**-.125**[b] 105 p=.001	**-.064**[b] 103 p=.010	**.003**[b] 105 p=.595	**.108** 105 p=.271	1.000			
Number of Replications	**-.000**[b] 105 p=.888	**.052**[b] 105 p=.064	**.067**[b] 103 p=.008	**.055**[b] 105 p=.015	**.014** 105 p=.888	**-.069** 105 p=.482	1.000		
RMSD	**.107**[b] 72 p=.005	**-.357**[b] 72 p=.001	**-.053**[b] 70 p=.055	**-.004**[b] 72 p=.613	**.113** 72 p=.344	**-.286** 72 p=.014	**.242** 72 p=.041	1.000	
Correlation	**-.019**[b] 94 p=.185	**.029**[b] 94 p=.263	**-.005**[b] 94 p=.495	**.093**[b] 94 p=.002	**-.042** 94 p=.689	**.428** 94 p=.001	**-.177** 94 p=.087	**-.492** 62 p=.001	1.000

a - Cramer V measure of association (based on Chi-square contingency table)
b - eta square (based on One-Way ANOVA with the qualitative values as the independent variable)
Note : The sign for the chi-square values and the one-way ANOVA values are obtained form the correlation matrix.
n = number of effect sizes used in computing a correlation

Table 4

Correlation Between Predictors and Criterion Variables for Item Difficulty Parameter

	Model	Method	Prior Dist.	Ability Dist.	Number of items	Number of exam.	Number of replica	RMSD	Correlation
IRT Model	1.000								
Method of Parameter Estimation	**-.560**[a] n = 119 p<.001	1.000							
Prior Distribution	**-.729**[a] 116 p<.001	**.484**[a] 116 p<.001	1.000						
Ability Distribution	**-.659**[a] 119 p<.001	**.448**[a] 119 p<.001	**.356**[a] 116 p<.001	1.000					
Number of Items	**.017**[b] 119 p=.375	**-.065**[b] 119 p=.020	**.001**[b] 116 p=.691	**.0.030**[b] 119 p=.059	1.000				
Number of Examinees	**.146**[b] 119 p=.001	**-.175**[b] 119 p=.001	**-.074**[b] 116 p=.003	**-.019**[b] 119 p=.130	**.089** 119 p=.334	1.000			
Number of Replications	**-.010**[b] 119 p=.573	**-.033**[b] 119 p=.146	**.001**[b] 116 p=.707	**.045**[b] 119 p=.020	**-.011** 119 p=.904	**-.092** 119 p=.321	1.000		
RMSD	**-.018**[b] 86 p=.468	**-.012**[b] 86 p=.610	**-.007**[b] 83 p=.460	**-.021**[b] 86 p=.188	**-.129** 86 p=.236	**-.291** 86 p=.006	**-.057** 86 p=.601	1.000	
Correlation	**-.038**[b] 106 p=.135	**.063**[b] 106 p=.035	**.013**[b] 106 p=.246	**.026**[b] 106 p=.097	**.010** 106 p=.916	**.104** 106 p=.287	**-.176** 106 p=.071	**-.978** 74 p=.001	1.000

a - Cramer V measure of association (based on Chi-square contingency table)
b - eta square (based on One-Way ANOVA with the qualitative values as the independent variable)
Note : The sign for the chi-square values and the one-way ANOVA values are obtained form the correlation matrix.
n = number of effect sizes used to compute a correlation

Table 5

Correlation Between Predictors and Criterion Variables for Ability Parameter

	Model	Method	Prior Dist.	Ability Dist.	Number of items	Number of exam.	Number of replica	RMSD	Correla-tion
IRT Model	1.000								
Method of Parameter Estimation	**-.460**[a] n = 66 p<.001	1.000							
Prior Distribution	**-.922**[a] 63 p<.001	**.460**[a] 63 p=.001	1.000						
Ability Distribution	**-.520**[a] 66 p<.001	**.230**[a] 66 p=.174	**.529**[a] 63 p<.001	1.000					
Number of Items	**-.085**[b] 66 p=.060	**.016**[b] 66 p=.600	**.054**[b] 63 p=.065	**.021**[b] 66 p=.249	1.000				
Number of Examinees	**.306**[b] 66 p=.001	**-.513**[b] 66 p=.001	**-.197**[b] 63 p=.003	**-.008**[b] 66 p=.481	**-.079** 66 p=.530	1.000			
Number of Replications	**-.118**[b] 66 p=.019	**-.026**[b] 66 p=.436	**.024**[b] 63 p=.223	**.016**[b] 66 p=.307	**.030** 66 p=.814	**-.297** 66 p=.015	1.000		
RMSD	**.265**[b] 65 p=.001	**-.189**[b] 65 p=.001	**-.176**[b] 62 p=.007	**-.031**[b] 65 p=.160	**-.692** 65 p=.001	**.526** 65 p=.001	**-.238** 65 p=.056	1.000	
Correlation	**-.185**[b] 53 p=.005	**.258**[b] 53 p=.006	**.144**[b] 53 p=.005	**.068**[b] 53 p=.060	**.658** 53 p=.001	**-.502** 53 p=.001	.[1]	**-.942** 53 p=.001	1.000

a - Cramer V measure of association (based on Chi-square contingency table)
b - eta square (based on One-Way ANOVA with the qualitative values as the independent variable)
Note : The sign for the chi-square values and the one-way ANOVA values are obtained form the correlation matrix.
n = number of effect sizes used to compute a correlation
1 No variation in the replication.

Table 6

Heterogeneity Summary Table for item discrimination with log RMSD as the outcome measure

Source	Statistics	Degrees of Freedom
Between Studies	**945.78**	**4**
Within Studies		
Study 1 (Hulin, et al., 1982)	n.s.	n.s.
Study 2 (Swaminathan & Gifford, 1986)	435.20	11
Study 3 (Yen, 1987)	35.21	17
Study 4 (Skaggs & Stevenson, 1989)	n.s.	n.s.
Study 5 (Gifford & Swaminathan, 1990)	166.92	6
Study 6 (Seong, 1990)	70.69	5
Study 7 (Harwell & Janosky, 1991)	188.01	29
Total within studies	**896.04**	**68**
Overall	**1841.82**	**72**

n.s. = not significant at $\alpha = .05$

Table 7

Heterogeneity Summary Table for item difficulty with log RMSD as the outcome measure

Source	Statistics	Degrees of Freedom
Between Studies	**165.17**	**4**
Within Studies		
Study 1 (Hulin, et al., 1982)	n.s.	n.s.
Study 2 (Swaminathan & Gifford, 1986)	85.31	11
Study 3 (Yen, 1987)	35.10	17
Study 4 (Skaggs & Stevenson, 1989)	n.s.	n.s.
Study 5 (Gifford & Swaminathan, 1990)	121.51	20
Study 6 (Seong, 1990)	63.35	5
Study 7 (Harwell & Janosky, 1991)	355.00	29
Total within studies	**660.28**	**82**
Overall	**825.45**	**86**

n.s. = not significant at $\alpha = .05$

Table 8

Heterogeneity Summary Table for ability parameter with log RMSD as the outcome measure

Source	Statistics	Degrees of Freedom
Between Studies	**881.07**	**3**
Within Studies		
Study 1 (Hulin, et al., 1982)	n.s.	n.s.
Study 2 (Swaminathan & Gifford, 1986)	31.65	11
Study 3 (Yen, 1987)	1186.51	26
Study 4 (Skaggs & Stevenson, 1989)	n.s.	n.s.
Study 5 (Gifford & Swaminathan, 1990)	128.19	20
Study 6 (Seong, 1990)	1.28	5
Study 7 (Harwell & Janosky, 1991)	n.s.	n.s.
Total within studies	**1347.62**	**62**
Overall	**2228.69**	**65**

n.s. = not significant at $\alpha = .05$

Table 9

Heterogeneity Summary Table for item discrimination with Fisher r-z correlation as the outcome measure

Source	Statistics	Degrees of Freedom
Between Studies	**227.04**	**5**
Within Studies		
Study 1 (Hulin, et al., 1982)	132.27	23
Study 2 (Swaminathan & Gifford, 1986)	11.95	11
Study 3 (Yen, 1987)	12.99	17
Study 4 (Skaggs & Stevenson, 1989)	8.72	7
Study 5 (Gifford & Swaminathan, 1990)	8.27	2
Study 6 (Seong, 1990)	n.s.	n.s.
Study 7 (Harwell & Janosky, 1991)	151.33	29
Total within studies	**325.51**	**89**
Overall	**552.55**	**94**

n.s. = not significant at $\alpha = .05$

Table 10

Heterogeneity Summary Table for item difficulty with Fisher r-z correlation as the outcome measure

Source	Statistics	Degrees of Freedom
Between Studies	**144.86**	**5**
Within Studies		
Study 1 (Hulin, et al., 1982)	497.77	23
Study 2 (Swaminathan & Gifford, 1986)	51.38	11
Study 3 (Yen, 1987)	114.47	17
Study 4 (Skaggs & Stevenson, 1989)	13.23	7
Study 5 (Gifford & Swaminathan, 1990)	104.34	14
Study 6 (Seong, 1990)	n.s.	n.s.
Study 7 (Harwell & Janosky, 1991)	123.44	29
Total within studies	**904.64**	**101**
Overall	**1049.50**	**106**

n.s. = not significant at $\alpha = .05$

Table 11

Heterogeneity Summary Table for ability parameter with Fisher r-z correlation as the outcome measure

Source	Statistics	Degrees of Freedom
Between Studies	**462.02**	**2**
Within Studies		
Study 1 (Hulin, et al., 1982)	n.s.	n.s.
Study 2 (Swaminathan & Gifford, 1986)	33.92	11
Study 3 (Yen, 1987)	1310.61	26
Study 4 (Skaggs & Stevenson, 1989)	.	n.s.
Study 5 (Gifford & Swaminathan, 1990)	83.80	14
Study 6 (Seong, 1990)	n.s.	n.s.
Study 7 (Harwell & Janosky, 1991)	n.s.	n.s.
Total within studies	**1428.32**	**51**
Overall	**1890.34**	**53**

n.s. = not significant at $\alpha = .05$

Table 12

Results for Weighted Least Squares Regression Analyses with log RMSD as Criterion Variable and Model, Method, Ability Distribution, N of Examinees, Items and Replication as Predictors

Parameter	R^2 of all predictors	R^2 of number of examinees and number of items, conditional on other predictors
Item Discrimination	.8050	.6219
Item Difficulty	.6598	.6316
Ability	.7750	.6340

Table 13

Results for Weighted Least Squares Regression Analyses with Fisher r-z correlation as Criterion Variable and Model, Method, Ability Distribution, N of Examinees, Items and Replications as Predictors

Parameter	R^2 of all predictors	R^2 of number of examinees and number of items, conditional on other predictors
Item Discrimination	.5110	.3393
Item Difficulty	.4288	.2257
Ability	.8219	.7585